ABSTRACT

To help ensure high data quality, data warehouses validate and clean, if needed incoming data tuples from external sources. In many situations, input tuples or portions of input tuples must match acceptable tuples in a reference table. For example, product name and description fields in a sales record from a distributor must match the pre-recorded name and description fields in a product reference relation. A disclosed system implements an efficient and accurate approximate or fuzzy match operation that can effectively clean an incoming tuple if it fails to match exactly with any of the multiple tuples in the reference relation. A disclosed similarity function that utilizes token substrings referred to as q-grams overcomes limitations of prior art similarity functions while efficiently performing a fuzzy match process.